

Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search

Year: 2020 | Citations: 565 | Authors: Jaehyeon Kim, Sungwon Kim, Jungil Kong, Sungroh Yoon

Abstract

Recently, text-to-speech (TTS) models such as FastSpeech and ParaNet have been proposed to generate mel-spectrograms from text in parallel. Despite the advantage, the parallel TTS models cannot be trained without guidance from autoregressive TTS models as their external aligners. In this work, we propose Glow-TTS, a flow-based generative model for parallel TTS that does not require any external aligner. By combining the properties of flows and dynamic programming, the proposed model searches for the most probable monotonic alignment between text and the latent representation of speech on its own. We demonstrate that enforcing hard monotonic alignments enables robust TTS, which generalizes to long utterances, and employing generative flows enables fast, diverse, and controllable speech synthesis. Glow-TTS obtains an order-of-magnitude speed-up over the autoregressive model, Tacotron 2, at synthesis with comparable speech quality. We further show that our model can be easily extended to a multi-speaker setting.